10

كشف التشوهات

الهدف في الكشف عن التشوهات هو إيجاد كائنات تختلف عن معظم الكائنات الأخرى. تُعرف الكائنات المشوّهة (anomalous) عادة **بالشواذ (outliers)** باعتبار أنها (على رسم بياني مبعثر للبيانات) تتوضع بعيداً عن نقاط بيانات أخرى. يُعرف كشف التشوّهات أيضاً **بكشف الانحراف (deviation detect)** لأن للكائنات المشوّهة قيم سمات تنحرف بشكل هام عن القيم المتوقعة أو المعتادة للسمات، أو **التنقيب عن الاستثناءات (exception mining)**، لأن التشوهات تكون استثنائية نوعاً ما. سنستخدم في هذا الفصل المصطلحين تشوّه (anomaly) أو شذوذ (outlier).

هناك أشكال متنوعة من طرق اكتشاف التشوهات من عدة مجالات، بما في ذلك الإحصاء والتعلم الآلي (machine learning) والتنقيب عن البيانات. يحاول الجميع تصوير فكرة أن كائنات البيانات المشوّهة (anomalous) هي غير اعتيادية أو أنها بطريقة ما غير متسقة مع الكائنات الأخرى. وعلى الرغم من أن الكائنات أو الأحداث غير الاعتيادية هي (بحسب تعريفها) نادرة نسبياً، فإن هذا لا يعني أنها لا تحدث بشكل متكرر على الإطلاق. فمثلاً، يمكن أن يقع حدث هو في الأصل "واحد من أصل ألف" ملايين المرات عند دراسة بلايين الأحداث.

إن معظم الأحداث والكائنات في العالم الفعلي، أو المجتمع البشري، أو نطاق مجموعات البيانات، هي بالتعريف مألوفة أو اعتيادية. إلا أننا على أية حال ندرك إمكانية الحصول على الكائنات غير المألوفة أو غير الاعتيادية. وهذا يشمل بشكل استثنائي الفصول الجافة أو الماطرة، أو الرياضيين المشهورين، أو قيمة سمة تكون أصغر بكثير أو أكبر بكثير من كافة القيم الأخرى. ينشأ اهتمامنا بالأحداث والكائنات الشاذة من حقيقة أنها تكون غالباً ذات أهمية غير عادية: الجفاف الذي يهدد المحاصيل، ومهارة استثنائية لرياضي قد تقود إلى تحقيق الفوز، والقيم الشاذة في نتائج تجريبية قد تشير إلى إما وجود مشكلة في التجربة أو إلى ظاهرة جديدة تجب دراستها.

توضح الأمثلة التالية تطبيقات تكون فيها التشوهات مهمة.

- **اكتشاف الاحتيال (Fraud Detection):** حيث يكون سلوك الشراء لشخص سرق بطاقة اعتماد مختلفاً نوعاً ما عن سلوك مالكها الأصلي. تحاول شركات بطاقات الاعتماد كشف السرقة بالبحث عن أنماط شراء تميز السرقة أو بملاحظة تغير عن السلوك الاعتيادي. تستخدم طرق مماثلة من أجل أنواع أخرى من الاحتيال.
- **الكشف عن التطفل (Intrusion Detection):** لسوء الحظ فإن من المعتاد وجود محاولات اختراق لأنظمة الحاسب وشبكات الحواسب. ففي حين أن بعض هذه الاختراقات تكون واضحة، كتلك المصممة لتعطيل أو إحداث مشكلات في الحواسب، فإن من الصعب اكتشاف محاولات اختراق أخرى، كتلك المصممة لجمع معلومات بشكل سري. يمكن اكتشاف العديد من محاولات الاختراق هذه فقط بمراقبة الأنظمة والشبكات بحثاً عن سلوكيات غير معتادة.
- **اضطراب النظام البيئي (Ecosystem Disturbance):** هناك في الواقع أحداث نموذجية يمكن أن يكون لها تأثير هام على البشر. تعتبر الأعاصير والفيضانات والجفاف والتيارات الحرارية والنيران أمثلة عن هذه الأحداث. والغاية غالباً هي التنبؤ بأرجحية وقوع هذه الحوادث وبمسبباتها.
- **الصحة العامة (Public Health):** تقدم المستشفيات والعيادات العامة في الكثير من الدول تقارير إحصائية مختلفة إلى مؤسسات وطنية لإجراء تحاليل إضافية عليها. فمثلاً، إذا تم تلقيح جميع الأطفال في مدينة ضد مرض معين (الحصبة مثلاً)، فإن وقوع بعض الإصابات في مستشفيات مختلفة في المدينة هو حدث شاذ قد يشير إلى مشكلة في برامج التلقيح في المدينة.
- **الدواء (Medicine):** فمن أجل مريض معين، يمكن أن تشير الأعراض أو نتائج الفحوصات غير الاعتيادية إلى احتمال وجود مشاكل صحية. قد يعتمد كون نتيجة فحص شاذة على أمور أخرى تتعلق بالمريض، كالعمر والجنس. علاوة على ذلك فإن تصنيف نتيجة على أنها شاذة أم لا يكون عرضة لدفع ثمن مقابل (اختبارات إضافية غير ضرورية إذا كان المريض معافى وأذى محتمل للمريض إذا تُرِكت حالة بدون تشخيص وعلاج).

على الرغم من أنه تم مؤخراً توجيه الاهتمام نحو الكشف عن التشوه من خلال تطبيقات تم فيها التركيز على التشوهات، فإنه قد تم على مر الوقت إظهار اكتشاف (وإزالة) التشوه كتقنية لتحسين تحليل كائنات البيانات العادية. فمثلاً، يمكن أن يحرِّف وجود قيم شاذة قليلة العدد نسبياً المتوسط (mean) والانحراف المعياري لمجموعة قيم أو تعديل مجموعة العناقيد الناتجة عن خوارزمية عنقدة. ولذلك فإن الكشف عن (وإزالة) التشوهات هو غالباً جزء من المعالجة المسبقة للبيانات.

سنركز في هذا الفصل على الكشف عن التشوهات. وسنقدم بعد بضعة خطوات تمهيدية شرحاً مفصلاً لبعض الطرق الهامة للكشف عن التشوهات، مع توضيحها من خلال أمثلة عن تقنيات معينة.

1.10 خطوات تمهيدية

سنقدم بعض المعلومات الأولية قبل أن نبدأ بمناقشة خوارزميات محددة للكشف عن التشوهات. وبشكل خاص فإننا (1) سنستكشف مسببات التشوهات، و (2) سندرس الطرق المختلفة للكشف عن التشوهات، و (3) سنحدد الفروقات فيما بين الطرق استناداً إلى ما إذا كانت تستخدم معلومات تسمية الصنف (class label)، و (4) سنتحدث عن مسائل شائعة تتعلق بتقنيات الكشف عن التشوهات.

1.1.10 أسباب الحصول على تشوهات

فيما يلي بعض الأسباب الشائعة للحصول على تشوهات وهي: بيانات من أصناف مختلفة، الاختلاف الطبيعي، وأخطاء قياس وجمع البيانات.

بيانات من أصناف مختلفة. قد يكون كائن مختلفاً عن كائنات أخرى (أي أنه تشوّه) لأنه من نوع أو صنف مختلف. للتوضيح نقول أن شخصاً يرتكب احتيالاً على بطاقات الائتمان ينتمي إلى صنف من مستخدمي بطاقات الائتمان يختلف عن أولئك الأشخاص الذين يستخدمون بطاقات الائتمان بشكل شرعي. تعتبر معظم الأمثلة الواردة في بداية هذا الفصل (وبالتحديد الاحتيال والتطفل وتفشي الأمراض والنتائج غير المعتادة للفحوصات) أمثلة عن التشوهات التي تمثل صنفاً مختلفاً من الكائنات. تعتبر تشوهات كهذه غالباً مهمة وينصب التركيز عليها من خلال الكشف عن التشوهات في مجال التنقيب عن البيانات.

يتم التعبير فكرة كون الكائنات المشوّهة تأتي من مصدر (صنف) يختلف عن معظم كائنات البيانات من خلال تعريف عام للشواذ وضعه Douglas Hawkins.

التعريف 1.10 (تعريف Hawkins للشواذ). المشاهدة الشاذة هي مشاهدة تختلف كثيراً عن المشاهدات الأخرى بشكل يدعو إلى الشك بأنه قد تم توليدها من خلال آلية مختلفة.

الاختلاف الطبيعي (Natural Variation). يمكن نمذجة الكثير من مجموعات البيانات من خلال توزيعات إحصائية، كالتوزيع الطبيعي (الغوصي)، حيث يتناقص احتمال كائن بيانات بشكل سريع بازدياد بُعد الكائن عن مركز التوزيع. يمكن التعبير عن ذلك بشكل مختلف بأن معظم الكائنات تكون قرب مركز (كائن وسطي) وتكون أرجحية أن يكون كائن مختلفاً بشكل هام عن الكائن الوسطي هذا صغيرة. فمثلاً، لا يكون شخص طويل بشكل استثنائي شاذاً إذا اعتبرنا أنه من صنف مختلف من الكائنات، ولكن فقط إذا اعتبرنا وجود قيمة متطرفة للميزة (الطول) المسيطرة وفقاً لكافة الكائنات. إن التشوهات التي تمثل أشكالاً مختلفة مفرطة أو غير واردة تكون غالباً مهمة.

أخطاء قياس وجمع البيانات. تعتبر أخطاء عملية جمع البيانات أو القياس مصدراً آخر للتشوهات. فمثلاً، يمكن تسجيل قياس بشكل خاطئ بسبب خطأ بشري، أو مشكلة في جهاز القياس، أو وجود تشويش (noise). الغاية هي التخلص من هذه التشوهات، باعتبار أنها تعطي معلومات غير ذات أهمية ولكنها فقط تخفض جودة البيانات وتحليل البيانات لاحقاً. تركز المعالجة السابقة للبيانات (وخاصة تصفية (cleaning) البيانات) على إزالة هذا النوع من التشوهات.

ملخص. يمكن أن يكون تشوه ما نتيجة للمسببات المعطاة آنفاً أو نتيجة لمسببات أخرى لم نذكرها. يمكن في الواقع أن تكون للتشوهات في مجموعة البيانات مصادر مختلفة، ويكون السبب المؤدي إلى نشوء تشوه ما غير معروف غالباً. تركز تقنيات الكشف عن التشوهات على إيجاد كائنات تختلف جوهرياً عن معظم الكائنات الأخرى، ولا تتأثر التقنيات نفسها بمصدر التشوه. وبالتالي فإن سبب التشوه يكون مهماً فقط فيما يتعلق بالتطبيق المقصود.

2.1.10 طرق الكشف عن التشوهات

نقدم هنا شرحاً لبعض تقنيات الكشف عن التشوهات وتعريفات التشوه المقترنة بها. هناك بعض التداخل بين هذه التقنيات، وسيتم في التمرين 1 سبر العلاقات فيما بينها.

التقنيات التي تستند إلى النموذج (Model-Based). تقوم كثير من تقنيات الكشف عن التشوه أولاً ببناء نموذج للبيانات. التشوهات هي كائنات لا تتلاءم مع النموذج بشكل جيد.

يمكن على سبيل المثال إنشاء نموذج لتوزيع البيانات من خلال استخدام البيانات لتقدير وسطاء (parameters) التوزيع الاحتمالي. يكون كائن غير متلائم مع والنموذج بشكل جيد (أي أنه تشوه) إذا لم يكن مرجّحاً جداً وفق التوزيع. فإذا كان النموذج مجموعة من العناقيد، فإن التشوه هو كائن لا ينتمي بقوة إلى أي عنقود. وعند استخدام نموذج انحدار (regression model)، فإن التشوه هو كائن يكون بعيداً نسبياً عن قيمته المتوقعة (predicted).

بما أنه يمكن إظهار الكائنات المشوهة والطبيعية على أنها تحدد صنفين متمايزين، فإن من الممكن استخدام تقنيات التصنيف لبناء نماذج لهذين الصنفين. يمكن بالطبع استخدام تقنيات التصنيف فقط إذا توفرت تسميات الأصناف من أجل بعض الكائنات بحيث يمكن بناء مجموعة تدريب (training set). كما أن التشوهات نادرة نسبياً، ويجب أخذ هذا الأمر بعين الاعتبار عند اختيار كل من تقنية التصنيف والمقاييس التي سيتم استخدامها للتقييم. (راجع المقطع 7.5).

من الصعب في بعض الحالات بناء نموذج، كأن يكون هذا لأن التوزيع الإحصائي للبيانات غير معروف أو لأنه لا تتوفر بيانات تدريب. يمكن في هذه الحالات استخدام تقنيات لا تتطلب بناء نموذج، كتلك التي سنتحدث عنها فيما يلي.

التقنيات التي تستند إلى القرابة (Proximity-Based). من الممكن غالباً تعريف مقياس قرابة بين الكائنات، ويستند عدد من طرق الكشف عن التشوه إلى قيم القرابة. الكائنات المشوهة هي تلك التي تكون متمايزة عن معظم الكائنات الأخرى. تستند كثير من التقنيات في هذا المجال إلى المسافات ويُشار إليها على أنها تقنيات كشف الشواذ استناداً إلى المسافة. يمكن الكشف بصرياً عن الشواذ استناداً إلى المسافة عندما يكون من الممكن عرض البيانات كمخطط بياني مبعثر ثنائي أو ثلاثي الأبعاد، وذلك بالبحث عن نقاط منفصلة عن معظم النقاط الأخرى.

التقنيات التي تستند إلى الكثافة (Density-Based). من الممكن حساب تقديرات لكثافة الكائنات بشكل مباشر تقريباً، وخاصة إذا توفر مقياس قرابة بين الكائنات. تكون الكائنات التي توجد في مناطق منخفضة الكثافة متمايزة نسبياً عن جيرانها، ويمكن اعتبارها تشوهات. تفرض طريقة أكثر تعقيداً حقيقة أنه يمكن أن تكون لمجموعات البيانات مناطق ذات كثافات مختلفة بشكل كبير، وتصنّف نقطة على أنها شاذة فقط إذا كانت لها كثافة محلية تكون أقل بشكل هام من معظم جيرانها.

3.1.10 استخدام تسميات الأصناف

هناك ثلاثة طرق أساسية لكشف التشوهات: غير المُراقبة (unsupervised)، والمُراقبة، وشبه المُراقبة (semi-supervised). الفارق الأساسي هو إلى أي حد تتوفر تسميات الأصناف (تشوه (anomaly) أو طبيعي (normal)) من أجل بعض البيانات على الأقل.

الكشف المُراقب عن التشوهات (Supervised anomaly detection). تتطلب تقنيات الكشف المُراقب عن التشوهات وجود مجموعة تدريب تتضمن كائنات مشوهة وطبيعية. (لاحظ أنه قد يكون هناك أكثر من صنف واحد طبيعي أو مشوّه). وكما ذكرنا سابقاً فإن تقنيات التصنيف التي تعالج مشكلة ما تُعرف بالصنف النادر (rare class) تكون مناسبة بشكل خاص لأن التشوهات نادرة نسبياً مقارنة بالكائنات الطبيعية. راجع المقطع 7.5.

الكشف غير المُراقب عن التشوهات (Unsupervised anomaly detection). لا تتوفر في الكثير من الحالات العملية تسميات الأصناف. والغاية في حالة كهذه هي إسناد درجات (score) (أو تسمية) إلى كل مثيل (instance) يعكس إلى أي درجة يكون المثيل تشوهاً. لاحظ أن وجود كثير من التشوهات التي تكون مشابهة لبعضها البعض قد يؤدي إلى تسميتها كلها على أنها طبيعية أو يكون لها درجات شذوذ منخفضة (low outlier score). وبالتالي فإنه لكي يكون الكشف غير المُراقب عن التشوهات ناجحاً، يجب أن تكون التشوهات متمايزة عن بعضها البعض، إضافة إلى تمايزها عن الكائنات الطبيعية.

الكشف شبه المُراقب عن التشوهات (Semi-supervised anomaly detection). تحوي بيانات التدريب أحياناً بيانات طبيعية لها تسمية (labled)، ولكن لا توجد معلومات عن الكائنات المشوهة. الغاية في الوضع شبه المُراقب هي إيجاد تسمية تشوّه أو درجة التشوه من أجل مجموعة من الكائنات المُعطاة باستخدام معلومات عن الكائنات الطبيعية المُسمّاة. لاحظ أنه في هذه الحالة يكون وجود كثير من الكائنات الشاذة المرتبطة ببعضها في مجموعة الكائنات التي سيتم إعطاؤها درجة شذوذ لا يؤثر على تقييم الشواذ. قد يكون من الصعب في بعض الحالات العملية إيجاد مجموعة صغيرة من الكائنات الطبيعية التي تمثّل المجموعة.

يمكن استخدام كافة مخططات الكشف عن التشوهات المشروحة في هذا الفصل في النمط المُراقَب أو غير المُراقَب. كما أن المخططات المُراقبة تماثل في جوهرها مخططات التصنيف من أجل الأصناف النادرة (rare class) المشروحة في المقطع 7.5.

4.1.10 مسائل هامة

هناك تشكيلة واسعة من المسائل الهامة التي تجب معالجتها عند التعامل مع التشوهات.

عدد السمات المستخدمة لتعريف تشوه. يستند السؤال حول ما إذا كائن تشوهاً إلى سمة واحدة هي سؤال عن ما إذا كانت قيمة تلك السمة للكائن شاذة. وبما أنه قد تكون للكائن عدة سمات، فقد تكون لبعض هذه السمات قيم شاذة، فيما تكون لسماته الأخرى قيم عادية. علاوة على ذلك فإن الكائن قد يكون تشوهاً حتى لو لم تكن أي من قيم سماته شاذة على انفراد. من الشائع على سبيل المثال أن يكون هناك أشخاص طولهم قَدمَين (أطفال) أو وزنهم 300 باوند، ولكن من غير الشائع أن يكون هناك شخص طوله قدمان ووزنه 300 باوند. يجب أن يحدد التعريف العام للتشوه كيف سيتم استخدام قيم سمات متعددة لتحديد ما إذا كان الكائن تشوّهاً أم لا. وهذه مسألة هامة بشكل خاص عندما تكون أبعاد البيانات عالية.

المنظور الشامل في مقابل المحلي. قد يبدو كائن غير اعتيادي بالنسبة لكافة الكائنات، ولكنه ليس كذلك بالنسبة لكائنات في جواره المحلي. يمكن على سبيل المثال أن يكون شخص طوله 6 أقدام و 5 إنشات طويلاً بشكل غير معتاد بالنسبة للمجتمع الإحصائي الكلي، ولكن ليس بالنسبة للاعبي كرة السلة المحترفين.

إلى أي درجة تكون نقطة تشوهاً. يتم تقييم ما إذا كان كائن تشوهاً بواسطة بعض التقنيات بطريقة ثنائية: الكائن إما تشوه أو ليس كذلك. لا يعكس هذا عادة الواقع الكامن بأن بعض الكائنات مفرطة في التشوه مقارنة ببعضها الآخر. وبذلك فإن من الضروري أن يكون لدينا تقييم لدرجة كون كائن هو تشوه. يُعرف هذا التقييم **بدرجات التشوه أو الشذوذ** (anomaly or outlier score).

تحديد تشوه واحد في كل مرة أم تحديد عدة تشوهات دفعة واحدة. يتم في بعض التقنيات إزالة التشوهات واحداً في كل مرة، أي أنه يتم تحديد المثيل الأكثر تشوهاً وإزالته ثم يتم تكرار العملية. أما في تقنيات أخرى فيتم تحديد تشكيلة من التشوهات معاً. غالباً ما تكون

التقنيات التي تحاول تحديد تشوه في كل مرة عرضة لمشكلة تُعرف **بالحجب** (masking)، حيث أن وجود عدة تشوهات يحجب وجود الكل. ومن ناحية أخرى فإن التقنيات التي تكشف عدة كائنات شاذة قد تواجه مشكلة **الغمر** (swamping)، حيث يتم تصنيف كائنات طبيعة على أنها شواذ. أما في الطرق التي تستند إلى النموذج (model-based)، فإن هذه التأثيرات يمكن أن تحدث لأن التشوهات تحرِّف نموذج البيانات.

التقييم. إذا توفرت تسميات الأصناف لتحديد البيانات المشوهة والطبيعية، فإن من الممكن تقييم فعالية مخطط كشف التشوهات باستخدام مقاييس أداء التصنيف المشروحة في المقطع 7.5. ولكن بما أن الصنف المشوه يكون في العادة أصغر بكثير من الصنف الطبيعي، فإن مقاييس مثل الاستدعاء (recall) والتحقيق (precision) ومعدل الإيجابية المضللة (false positive error) ستكون أكثر ملاءمة من الدقة (accuracy). أما إذا كانت تسميات الأصناف غير متوفرة، فإن من الممكن الحكم على فعالية الكشف عن الشواذ من خلال التحسن في النموذج بمجرد إزالة التشوهات.

الفعالية. هناك فروقات هامة في الكلفة الحسابية للمخططات المختلفة للكشف عن التشوهات. يمكن أن تتطلب المخططات التي تستند إلى التصنيف مصادر كثيرة لإنشاء نموذج التصنيف، ولكن تطبيقها لا يكون مكلفاً عادة. وبشكل مشابه تنشئ الطرق الإحصائية نموذجاً إحصائياً ويمكنها بعدها تحديد فئة كائن خلال زمن ثابت. التعقيد الزمني للطرق التي تستند إلى القرابة هو $O(m)^2$، حيث m هو عدد الكائنات، لأن من الممكن الحصول على المعلومات التي تحتاجها فقط بحساب مصفوفة القرابة (proximity matrix). يمكن تخفيض هذا التعقيد الزمني في حالات خاصة، كأن تكون البيانات ثنائية الأبعاد، وذلك باستخدام بنية بيانات وخوارزميات خاصة. سنتحدث في التمرين 3 عن التعقيد الزمني للطرق الأخرى.

خريطة الطريق

تشرح المقاطع الأربعة القادمة عدة فئات رئيسة لطرق الكشف عن التشوهات: الإحصائية، واستناداً إلى القرابة، واستناداً إلى الكثافة، واستناداً إلى العناقيد. سندرس تقنية واحدة أو أكثر ضمن كل من هذه الفئات. سنتبع في هذه المقاطع الخبرة ونستخدم المصطلح شاذ بدلاً من تشوه.

2.10 الطرق الإحصائية

الطرق الإحصائية هي طرق تستند إلى النموذج، أي أنه يتم بناء نموذج للبيانات، ثم يتم تقييم الكائنات وفقاً لمدى ملاءمتها للنموذج. تستند معظم الطرق الإحصائية للكشف عن الشواذ إلى بناء نموذج توزيع احتمالي ودراسة أرجحية خضوع الكائنات لذلك النموذج. يعبر التعريف 2.10 عن هذه الفكرة.

التعريف 2.10 (التعريف الإحصائي للكائن الشاذ). الكائن الشاذ هو كائن ذو احتمال ضعيف بالنسبة لنموذج التوزيع الاحتمالي للبيانات.

يتم إنشاء نموذج توزيع احتمالي من البيانات وذلك بتقدير وسطاء توزيع يُعرِّفه المستخدم. فإذا افترضنا أن البيانات خاضعة لتوزيع غوصي، فإن من الممكن تقدير المتوسط (maen) والانحراف المعياري لهذا التوزيع بحساب متوسط والانحراف المعياري للبيانات. يمكن بعدها تقدير احتمال كل كائن يخضع لهذا التوزيع.

تم اشتقاق تشكيلة واسعة من الاختبارات الإحصائية استناداً إلى التعريف 2.10 بهدف الكشف عن الشواذ، أو ما يُعرف **بالمُشاهدات المتنافرة** (discordant observations) في أدبيات الإحصاء. تكون كثير من اختبارات التنافر هذه متخصصة وتفترض وجود مستوى معين من المعرفة الإحصائية تتجاوز نطاق هذا الكتاب. ولهذا فإننا سنوضح الأفكار الأساسية مع بعض الأمثلة فقط، ونترك الباقي للقارئ.

مسائل هامة

فيما يلي بعض المسائل الهامة التي تواجه هذه الطريقة في كشف الشواذ:

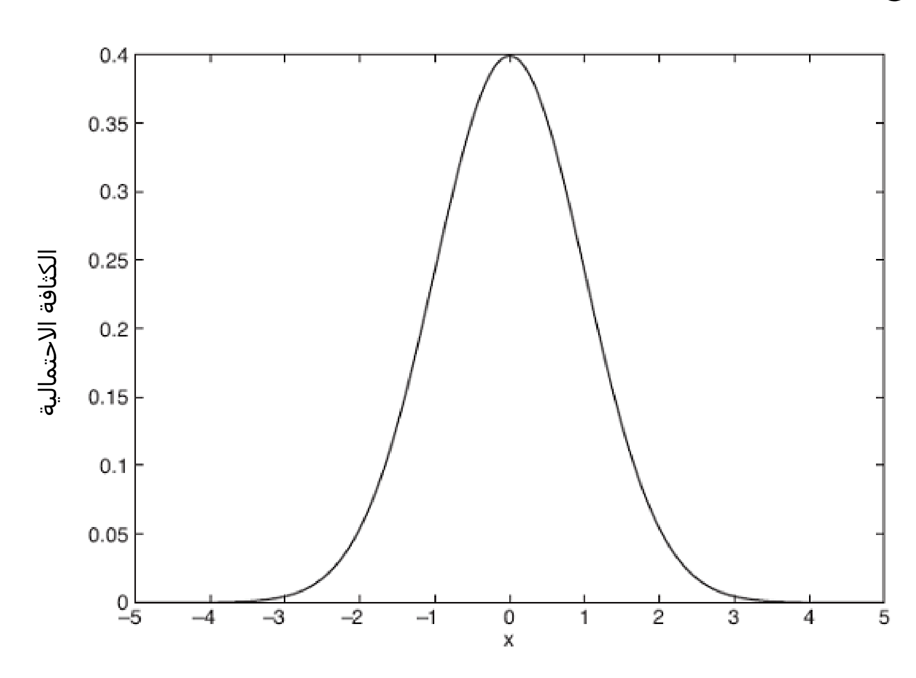
تحديد توزيع مجموعة بيانات. ففي حين يمكن توصيف الكثير من أنواع البيانات من خلال عدد صغير من التوزيعات الشائعة، كالتوزيع الغوصي (Gaussian)، وتوزيع بواسون (Poisson)، أو ثنائي الحد (binominal)، فإن من الشائع نسبياً أن تكون هناك مجموعات بيانات تخضع لتوزيعات غير معيارية. فإذا تم بالطبع اختيار نموذج خاطئ فإن من الممكن أن يتم بشكل خاطئ تعريف كائن على أنه شاذ. يمكن على سبيل المثال أن تتم نمذجة البيانات على أنها تأتي من توزيع غوصي، ولكن من الممكن أن تكون في الواقع من توزيع يكون هناك احتمال أعلى (مقارنة بالتوزيع الغوصي) أن توجد فيه قيم بعيدة جداً عن المتوسط. من الشائع عملياً وجود توزيعات إحصائية من هذا النوع وهي تُعرف **بالتوزيعات ثقيلة الذيل** (heavy-tailed distributions).

عدد السمات المستخدمة. يتم تطبيق معظم تقنيات الكشف عن الشواذ إحصائياً على سمة واحدة، ولكنه قد تم تعريف بعض التقنيات من أجل بيانات متعددة المتحولات (multivariate).

خليط من التوزيعات. يمكن نمذجة البيانات كخليط من التوزيعات، ويمكن تطوير مخططات كشف عن الشواذ استناداً إلى نماذج كهذه. وعلى الرغم من أنها أكثر قوة، فإن هذه النماذج تكون أكثر تعقيداً، إذ أنها تكون صعبة الفهم والاستخدام معاً. من الضروري على سبيل المثال تعريف التوزيعات قبل أن نتمكن من تصنيف الكائنات على أنها شواذ. راجع شرح النماذج المختلطة وخوارزمية EM الواردة في المقطع 2.2.9.

1.2.10 الكشف عن الشواذ في توزيعات طبيعية وحيدة المتغير

يعتبر التوزيع الغوصي (الطبيعي) أحد أكثر التوزيعات استخداماً في الإحصاء، وسنستخدمه لشرح طريقة بسيطة للكشف عن الشواذ إحصائياً. يمتلك هذا التوزيع وسيطين هما μ (المتوسط) و σ (الانحراف المعياري)، ويتم تمثيله باستخدام الصيغة $N(\mu,\sigma)$. يعرض الشكل 1.10 تابع الكثافة لـ $N(0, 1)$.



الشكل 10.1 تابع الكثافة الاحتمالية لتوزيع غوصي بمتوسط 0 وانحراف معياري 1

هناك فرصة ضئيلة أن يقع كائن (قيمة) من التوزيع $N(0, 1)$ في زيلي (tail) التوزيع. هناك مثلاً احتمال قدره 0.0027 أن يقع كائن بعد المنطقة المركزية بين ±3 من الانحرافات المعيارية. وبشكل أعم نقول أنه إذا كان c ثابت و x هي قيمة سمة الكائن، فإن احتمال أن يكون $|x| \geq c$ يتناقص بشكل سريع عندما تزداد قيمة الثابت c. لتكن $\alpha = prob(|x| \geq c)$. يعرض الجدول 1.10 بعض الأمثلة عن قيم c وقيم α المقابلة لها عندما يكون التوزيع هو $N(0,1)$. لاحظ أن القيمة التي تكون أكبر من 4 انحرافات معيارية من المتوسط يكون احتمال وقوعها واحد في العشرة آلاف.

الجدول 1.10 أمثلة عن أزواج (c, α)، حيث $\alpha = prob(|x| \geq c)$، من أجل توزيع غوصي بمتوسط 0 وانحراف معياري 1

c	α من أجل $N(0,1)$
1.00	0.3173
1.50	0.1336
2.00	0.0455
2.50	0.0124
3.00	0.0027
3.50	0.0005
4.00	0.0001

وبما أن بُعد القيمة c هن مركز التوزيع $N(0,1)$ يتعلق مباشرة باحتمال القيمة، فمن الممكن استخدامه كأساس لاختبار ما إذا كان كائن (قيمة) شاذاً أم لا كما سنبين في التعريف 3.10.

التعريف 3.10 (الكائن الشاذ من أجل سمة وحيدة تخضع للتوزيع الغوصي $N(0,1)$). يكون كائن له قيمة سمة x تخضع للتوزيع الغوصي بمتوسط 0 وانحراف معياري 1 شاذاً إذا كانت:

$$|x| \geq c \qquad (1.10)$$

حيث c ثابت يتم اختياره بحيث يكون $prob(|x| \geq c) = \alpha$.

من الضروري لاستخدام هذا التعريف تحديد قيمة لـ α. فمن منظور كون القيم (الكائنات) غير الاعتيادية تشير إلى قيمة من توزيع مختلف، فإن α تشير إلى احتمال أن نقوم بشكل خاطئ بتصنيف قيمة من التوزيع المُعطى على أنها شاذة. أما من منظور كون القيمة الشاذة هي قيمة نادرة من التوزيع $N(0,1)$، فإن α تحدد درجة النُدرة.

إذا كان توزيع سمة تتم دراستها (من أجل الكائنات الطبيعية) توزيع غوصي بمتوسط μ وانحراف معياري σ (أي أنه توزيع $N(\mu,\sigma)$)، فإننا سنحتاج لكي نتمكن من استخدام التعريف 3.10 إلى تحوير (transform) السمة x لتصبح سمة جديدة z، لها توزيع $N(0,1)$. وبشكل أدق فإن الطريقة هي وضع $z = (x-\mu)/\sigma$. (تُدعى z بقيمة أو درجات z (z score)). وعلى أية حال فإن μ و σ مجهولين عادة ويتم تقديرهما باستخدام متوسط العينة $\bar{x}$ والانحراف المعياري للعينة s_x. يعمل هذا الأمر بشكل جيد في الحالة العملية عندما يكون عدد المُشاهدات كبيراً. نلاحظ على أية حال أن توزيع z ليس فعلياً $N(0, 1)$. سنتحدث عن إجرائية إحصائية أكثر تعقيداً (اختبار Grubbs) في التمرين 7.

2.2.10 الكائنات الشاذة في التوزيعات الطبيعية متعددة المتحولات

سنرغب من أجل المُشاهدات الغوصية متعددة المتحولات بأخذ طريقة تشبه تلك المُعطاة من أجل توزيع غوصي أحادي المتحول. سنرغب عملياً بتصنيف نقاط على أنها شاذة إذا كان لها احتمال منخفض بالنسبة للتوزيع المُقدّر للبيانات. علاوة على ذلك فإننا سنرغب بالحكم على هذا من خلال اختبار بسيط، كأن يكون مثلاً بُعد النقطة عن مركز التوزيع.

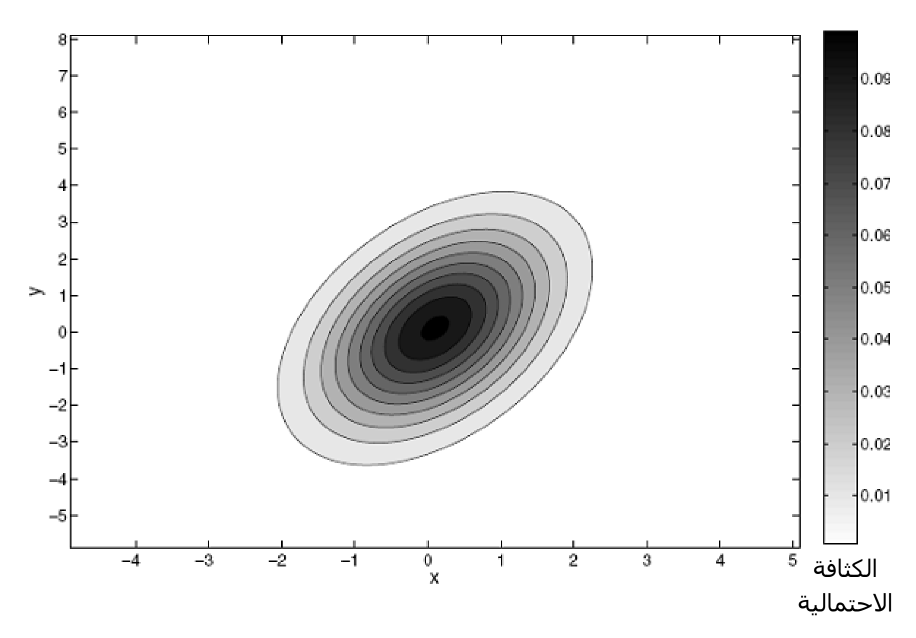
وبسبب الارتباط (correlation) بين المتحولات المختلفة (السمات)، فإن التوزيع الطبيعي متعدد المتحولات لا يكون متناظراً (symmetrical) بالنسبة لمركزه. يعرض الشكل 2.10 الكثافة الاحتمالية لتوزيع غوصي متعدد المتحولات ثنائي الأبعاد له متوسط (0 ,0) ومصفوفة تباين مشترك (covariance matrix) هي:

$$\sum = \begin{pmatrix} 1.00 & 0.75 \\ 0.75 & 3.00 \end{pmatrix}$$

فإذا كنا سنستخدم حد عتبة (threshold) بسيط لتحديد ما إذا كان كائن شاذاً، فإننا سنحتاج إلى مقياس مسافة يأخذ بعين الاعتبار شكل توزيع البيانات. تعتبر مسافة Mahalanobis مثالاً عن هذا المقياس. راجع المعادلة 14.2. تُعطي المعادلة 2.10 مسافة Mahalanobis بين نقطة x وبين متوسط البيانات $\bar{x}$.

$$mahalanobis(\mathbf{x}, \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})^T \qquad (2.10)$$

حيث **S** هي مصفوفة التباين المشترك للبيانات.

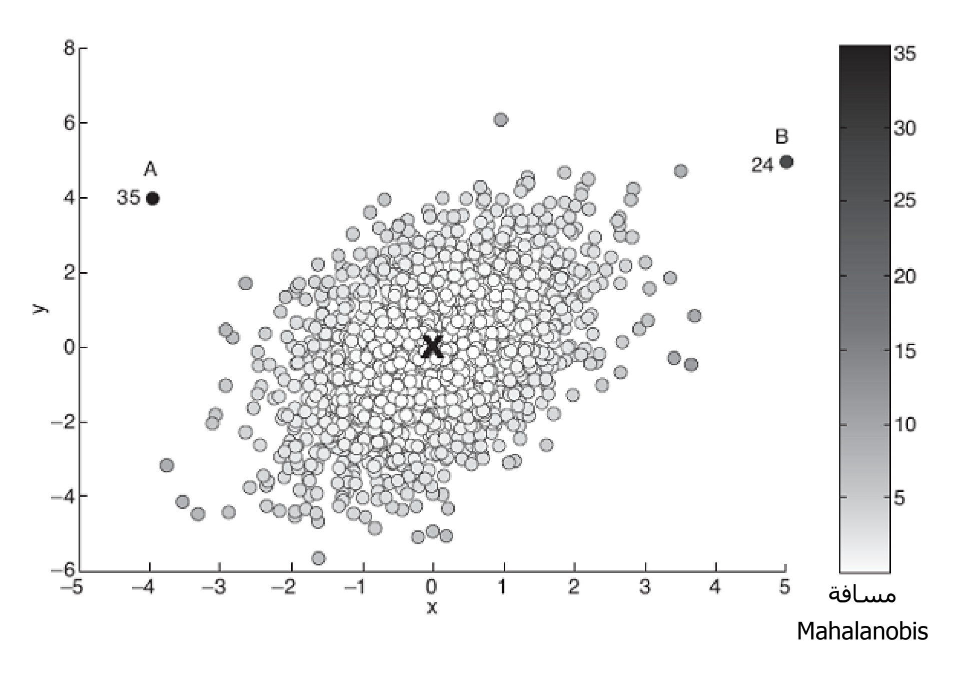


الشكل 2.10 الكثافة الاحتمالية لتوزيع غوصي تم استخدامه لتوليد نقاط الشكل 3.10

من السهل أن نبين أن مسافة Mahalanobis بين نقطة وبين متوسط التوزيع المقابل ترتبط مباشرة باحتمال تلك النقطة. إن مسافة Mahalanobis تساوي لوغاريتم (log) الكثافة الاحتمالية للنقطة مضافاً إليها ثابت. راجع التمرين 5.

المثال 1.10 (الكائنات الشاذة في التوزيعات الطبيعية متعددة المتحولات). يعرض الشكل 3.10 مسافة Mahalanobis (من متوسط التوزيع) من أجل نقاط في مجموعة بيانات ثنائية الأبعاد. إن النقطتين A(-4,4) و B(5,5) هما نقطتان شاذتان تمت إضافتهما إلى مجموعة البيانات، وقمنا بالإشارة إلى مسافة Mahalanobis لهما في الشكل. أما باقي نقاط مجموعة البيانات البالغ عددها 2000 نقطة فقد تم توليدها باستخدام التوزيع المستخدم في الشكل 2.10.

إن لكل من A و B مسافة Mahalanobis كبيرة. وعلى أية حال، وبالرغم من أن A أقرب إلى المركز (أشرنا إليه بـ x كبيرة بخط أسود عريض عند (0,0)) إذا اعتبرنا أننا نأخذ المسافة الإقليدية، فإنها أبعد من B وفق مسافة Mahalanobis لأن مسافة Mahalanobis تأخذ شكل التوزيع بعين الاعتبار. تمتلك النقطة B مسافة إقليدية هي $5\sqrt{2}$ ومسافة Mahalanobis هي 24، في حين أن للنقطة A مسافة إقليدية هي $4\sqrt{2}$ ومسافة Mahalanobis هي 35. ■



الشكل 3.10 مسافة Mahalanobis للنقاط من مركز مجموعة نقاط عددها 2002 ثنائية الأبعاد

3.2.10 طريقة النموذج المختلط للكشف عن التشوه

يقدم هذا المقطع تقنية كشف عن التشوه تستخدم طريقة النموذج المختلط. ففي العنقدة (راجع المقطع 2.2.9)، تفترض طريقة النموذج المختلط أن البيانات تأتي من خليط من التوزيعات الاحتمالية. وبشكل مشابه، تتم من أجل الكشف عن التشوه نمذجة البيانات على أنها خليط من توزيعين، واحد من أجل بيانات عادية وواحد من أجل البيانات الشاذة.

الغاية في كلتا حالتي العنقدة والكشف عن التشوهات هي تقدير وسطاء التوزيعات بهدف تكبير (maximize) الأرجحية الإجمالية (الاحتمال) للبيانات. ففي العنقدة، تُستخدم خوارزمية EM لتقدير وسطاء كل توزيع احتمالي. تستخدم تقنية الكشف عن التشوه الواردة هنا طريقة بسيطة. يمكن بشكل مبدئي وضع كافة الكائنات في مجموعة من الكائنات الطبيعية ومجموعة الكائنات المشوّهة. تقوم بعدها إجرائية تكرارية بتحويل الكائنات من المجموعة العادية إلى المجموعة الشاذة طالما أن التكرار يزيد الأرجحية الإجمالية للبيانات.

لنفترض أن مجموعة البيانات D تحوي كائنات من خليط من توزيعين احتماليين: M هو توزيع أغلبية الكائنات (الطبيعية)، و A هو توزيع الكائنات الشاذة. يمكن كتابة التوزيع الاحتمالي الإجمالي للبيانات بالشكل:

$$D(\mathbf{x}) = (1 - \lambda)M(\mathbf{x}) + \lambda A(\mathbf{x}) \qquad (3.10)$$

حيث $\mathbf{x}$ كائن و λ عدد بين 0 و 1 يُعطي نسبة الشواذ المتوقعة. يتم تقدير التوزيع M من البيانات، في حين أن التوزيع A يكون عادة منتظماً (uniform). لتكن M_t و A_t مجموعة الكائنات الطبيعية والشاذة (على الترتيب) في الزمن t. ففي الزمن $t = 0$، تكون $M_0 = D$ و A_0 خالية. ففي زمن اعتباطي t تكون أرجحية ولوغاريتم (log) الأرجحية لمجموعة البيانات الكلية D مُعطاة بالمعادلتين التاليتين على الترتيب:

$$L_t(D) = \prod_{\mathbf{x}_i \in D} P_D(\mathbf{x}_i) = \left((1-\lambda)^{|M_t|} \prod_{\mathbf{x}_i \in M_t} P_{M_t}(\mathbf{x}_i) \right) \left(\lambda^{|A_t|} \prod_{\mathbf{x}_i \in A_t} P_{A_t}(\mathbf{x}_i) \right) \qquad (4.10)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{\mathbf{x}_i \in M_t} \log P_{M_t}(\mathbf{x}_i) + |A_t| \log(\lambda) + \sum_{\mathbf{x}_i \in A_t} \log P_{A_t}(\mathbf{x}_i) \qquad (5.10)$$

حيث P_D و P_{M_t} و P_{A_t} هي توابع التوزيع الاحتمالي من أجل D و M_t و A_t على الترتيب. يمكن اشتقاق هذه المعادلة من التعريف العام للنموذج المختلط المُعطاة في المعادلة 6.9 (المقطع 2.2.9). من الضروري للقيام بذلك وضع الافتراض المبسِّط بأن الاحتمال هو 0 من أجل كائن في الحالتين التاليتين: (1) كائن موجود في A في حين أنه كائن طبيعي، و (2) كائن في M في حين أنه شاذ. التفاصيل مُعطاة في الخوارزمية 1.10.

بما أن عدد الكائنات الطبيعية كبير مقارنة بعدد الكائنات الشاذة، فإن توزيع الكائنات الطبيعية قد لا يتغير كثيراً عند نقل كائن إلى مجموعة الشواذ. وفي هذه الحالة فإن إسهام كل كائن طبيعي في الأرجحية الإجمالية للكائنات الطبيعية سيبقى ثابتاً نسبياً. علاوة على ذلك فإنه إذا افترضنا أن الكائنات الشاذة تخضع للتوزيع المنتظم، فإن كل كائن يتم نقله إلى مجموعة الشواذ يُسهم بمقدار ثابت في أرجحية الشواذ. وبالتالي فإن التغير الإجمالي في الأرجحية الإجمالية للبيانات عند نقل كائن إلى مجموعة الشواذ يساوي تقريباً احتمال الكائن وفق توزيع منتظم (مثقّل بـ λ)

مطروحاً منه احتمال الكائن وفق التوزيع الطبيعي لنقاط البيانات (مُثقّل بـ λ-1). وبالتالي فإن مجموعة الشواذ ستكون مؤلفة من تلك الكائنات التي لها احتمال عالٍ نسبياً وفق التوزيع المنتظم مقارنة باحتمالها وفق التوزيع الطبيعي للكائنات.

الخوارزمية 1.10 الكشف عن الشواذ استناداً إلى الأرجحية

Initialization: At time $t = 0$, let M_t contain all the objects, while A_t is empty.
Let $LL_t(D) = LL(M_t) + LL(A_t)$ be the log likelihood of all the data.
for each point **x** that belongs to M_t **do**
    Move **x** from M_t to A_t to produce the new data sets A_{t+1} and M_{t+1}.
    Compute the new log likelihood of D, $LL_{t+1}(D) = LL(M_{t+1}) + LL(A_{t+1})$
    Compute the difference, $\Delta = LL_t(D) - LL_{t+1}(D)$
    if $\Delta > c$, where c is some threshold **then**
        x is classified as an anomaly, i.e., M_{t+1} and A_{t+1} are unchanged and become the current normal and anomaly sets.
    end if
end for

إن الطريقة الواردة في الخوارزمية 1.10 في الحالات التي تحدثنا عنها للتو تُكافئ تقريباً تصنيف الكائنات التي لها احتمال منخفض وفق توزيع الكائنات الطبيعية على أنها شاذة. فمثلاً، عند تطبيق هذه التقنية على النقاط الواردة في الشكل 3.10 سيتم تصنيف النقطتين A و B (ونقاط أخرى بعيدة عن المتوسط) على أنها شاذة. وعلى أية حال، إذا تغير توزيع الكائنات الطبيعية بشكل ملحوظ عند إزالة الشواذ أو كان من الممكن نمذجة توزيع الشواذ بطريقة أكثر تعقيداً، فإن النتائج التي تُعطيها هذه الطريقة ستكون مختلفة عن نتائج التصنيف البسيط للكائنات ذات الاحتمال المنخفض على أنها شواذ. كما يمكن أن تعمل هذه الطريقة حتى عندما يكون توزيع الكائنات متعدد المنوال (multimodal).

4.2.10 أوجه القوة والضعف

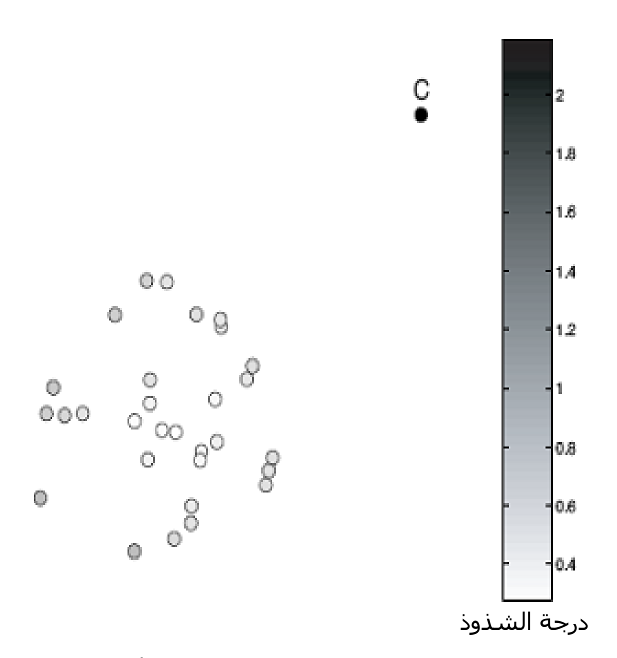
تمتلك الطرق الإحصائية للكشف عن الشواذ أساساً ثابتاً وتُبنى على أساس تقنيات إحصائية معيارية، مثل تقدير وسطاء توزيع. عندما تكون لدينا معرفة كافية بالبيانات ونوع الاختبار الذي يجب تطبيقه فإن هذه الاختبارات يمكن أن تكون فعالة جداً. هناك تشكيلة واسعة من

الاختبارات الإحصائية الخاصة بالكشف عن الشواذ في حالة سمات وحيدة. تتوفر بضعة خيارات من أجل بيانات متعددة المتحولات، ويكون إنجاز هذه الاختبارات ضعيفاً من أجل بيانات ذات أبعاد عالية.

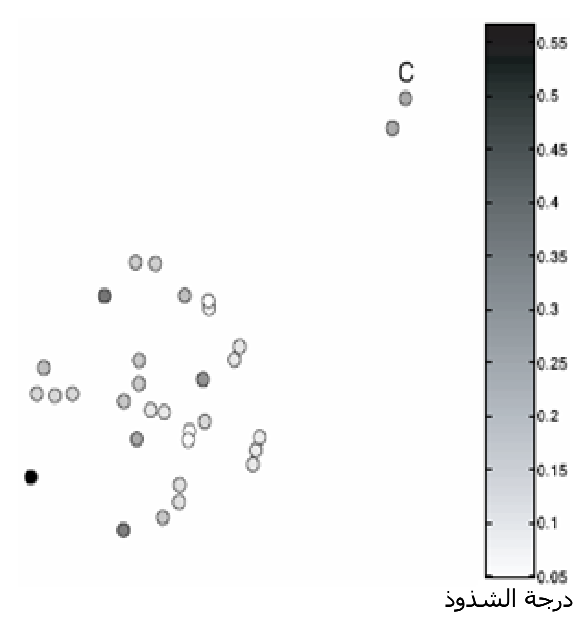
3.10 الكشف عن التشوهات استناداً إلى القرابة

على الرغم من وجود أشكال مختلفة عديدة من فكرة الكشف عن التشوهات استناداً إلى القرابة، فإن الفكرة الأساسية بسيطة ومباشرة. يكون كائن ما تشوّهاً إذا كان بعيداً عن معظم البيانات. إن هذه الطريقة أكثر عمومية وأسهل تطبيقاً مقارنة بالطرق الإحصائية، باعتبار أن من الأسهل تحديد مقياس قرابة ذو دلالة من أجل مجموعة بيانات مقارنة بتحديد توزيعها الاحتمالي.

من أسهل طرق قياس ما إذا كان كائن بعيداً عن معظم النقاط هي استخدام المسافة إلى الـ *k* جار الأقرب (*k*-nearest neighbor). يشرح التعريف 4.10 ذلك. أخفض قيمة لدرجة شذوذ (score) الكائن هي 0، في حين أن أعلى قيمة هي القيمة الأعظمية الممكنة لتابع المسافة، وهي عادة لانهاية).



الشكل 4.10 درجة الشذوذ استناداً إلى المسافة إلى الجار الأقرب الخامس

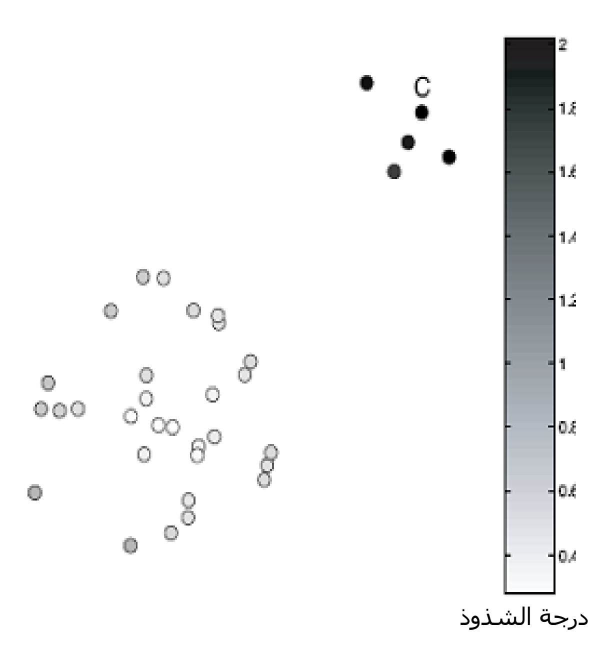


الشكل 5.10 درجة الشذوذ استناداً إلى المسافة إلى أول جار أقرب. تكون للكائنات الشاذة المجاورة درجات شذوذ منخفضة

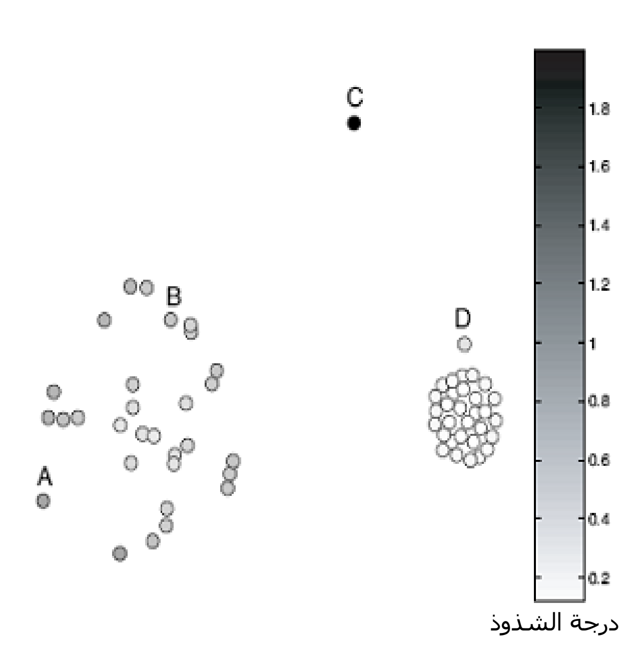
التعريف 4.10 (المسافة إلى الـ *k* جار الأقرب). تُعطى درجة شذوذ كائن بأنها المسافة إلى أقرب *k* جار له.

يعرض الشكل 4.10 مجموعة نقاط ثنائية الأبعاد. يشير تظليل كل نقطة إلى درجات الشذوذ لها باستخدام القيمة $k = 5$. لاحظ أنه تم بشكل صحيح إسناد درجة شذوذ مرتفعة إلى نقطة شاذة C.

يمكن أن تكون درجات الشذوذ حساسة بشكل كبير لقيمة *k*. فإذا كانت *k* صغيرة (1 مثلاً) فإن وجود عدد صغير من الكائنات الشاذة المجاورة سيؤدي إلى الحصول على درجات شذوذ منخفضة. يعرض الشكل 5.10 على سبيل المثال مجموعة من النقاط ثنائية الأبعاد كانت فيها نقطة أخرى قريبة من C. يعكس التظليل درجات الشذوذ باستخدام قيمة k=1. لاحظ أن لكل من C وجارها درجة شذوذ منخفضة. إذا كانت *k* كبيرة جداً، وعندها من الممكن أن تصبح كافة الكائنات في عنقود فيه كائنات أقل من *k* شواذاً. يبين الشكل 6.10 على سبيل المثال مجموعة بيانات ثنائية الأبعاد فيها عنقود طبيعي حجمه 5 بالإضافة إلى عنقود أكبر حجمه 30. فمن أجل $k = 5$ تكون درجات الشذوذ لكافة النقاط في العنقود الأصغر مرتفعة جداً. ولكي نجعل المخطط أكثر قوة في اختيار *k* يمكننا تعديل التعريف 4.10 بحيث يستخدم المتوسط الحسابي للمسافات إلى أقرب *k* جار أقرب.



الشكل 6.10 درجة الشذوذ استناداً إلى المسافة إلى الجار الأقرب الخامس. يصبح العنقود الصغير شاذاً



الشكل 7.10 درجة الشذوذ استناداً إلى المسافة إلى الجار الأقرب الخامس. عناقيد ذات كثافات مختلفة

1.3.10 أوجه القوة والضعف

إن مخططات الكشف عن الشواذ استناداً إلى المسافة المشروحة أعلاه وغيرها من المخططات ذات الصلة تكون بسيطة. تستغرق الطرق التي تستند إلى القرابة عادة زمناً قدره $O(m^2)$. يمكن أن يكون هذا مكلفاً جداً في حالة مجموعات بيانات ضخمة، على الرغم من أنه يمكن استخدام خوارزميات مخصصة لتحسين الأداء في حالة البيانات منخفضة الأبعاد. كما أن الطريقة حساسة لاختيار الوسطاء. علاوة على أنها لا يمكنها معالجة مجموعات بيانات فيها مناطق ذات كثافات مختلفة جداً لأنها تستخدم حدود عتبة شاملة لا يمكنها أن تأخذ بعين الاعتبار تغيرات كثافة كهذه.

لتوضيح ذلك، لتكن لدينا مجموعة النقاط ثنائية الأبعاد الواردة في الشكل 7.10. يعرض هذا الشكل عنقوداً من النقاط مفككاً نوعاً ما، وعنقوداً آخر من النقاط كثيف، ونقطتان C و D بعيدتان جداً عن هذين العنقودين. إن إسناد درجة الشذوذ إلى النقاط وفقاً للتعريف 4.10 من أجل $k = 5$ يؤدي إلى تعريف صحيح للنقطة C على أنها شاذة، ولكنه يعطي درجة شذوذ منخفضة من أجل النقطة D. وفي الواقع فإن درجة الشذوذ من أجل D أقل بكثير من الكثير من النقاط التي هي جزء من العنقود المفكك.

4.10 الكشف عن الشذوذ استناداً إلى الكثافة

إن الشواذ من وجهة النظر التي تستند إلى الكثافة هي الكائنات التي تقع في مناطق منخفضة الكثافة.

التعريف 5.10 (الكائن الشاذ استناداً إلى الكثافة). إن درجة شذوذ كائن هي مقلوب الكثافة حول الكائن.

يرتبط الكشف عن الشواذ استناداً إلى الكثافة بقوة بالكشف عن الشواذ استناداً إلى القرابة باعتبار أن الكثافة تُعرّف عادة من خلال القرابة. إحدى الطرق الشائعة هي تحديد الكثافة على أنها مقلوب المتوسط الحسابي للمسافة إلى الـ k جار الأقرب. فإذا كانت هذه المسافة صغيرة، فإن الكثافة تكون عالية، وبالعكس. وهذا ما يبينه التعريف 6.10.

التعريف 6.10 (مقلوب المسافة).

$$density(\mathbf{x},k)=\left(\frac{\sum_{\mathbf{y}\in N(\mathbf{x},k)} distance(\mathbf{x},\mathbf{y})}{|N(\mathbf{x},k)|}\right)^{-1} \quad (6.10)$$

حيث $N(\mathbf{x}, k)$ هو المجموعة التي تحوي الـ k جار الأقرب لـ $\mathbf{x}$ ، و $|N(\mathbf{x}, k)|$ هو حجم تلك المجموعة، و $\mathbf{y}$ هو جار أقرب.

هناك تعريف آخر للكثافة تستخدمه خوارزمية العنقدة DBSCAN. راجع المقطع 4.8.

التعريف 7.10 (عدّ النقاط ضمن نصف قطر مُعطى). إن الكثافة حول كائن تساوي عدد الكائنات التي تقع ضمن مسافة محددة d من الكائن.

يجب اختيار الوسيط d بعناية. فإذا كان d صغيراً جداً فقد تكون للكثير من النقاط الطبيعية كثافة منخفضة وبالتالي درجة شذوذ عالية. أما إذا كانت d مرتفعة فقد تكون للكثير من الشواذ كثافات (ودرجة شذوذ) تماثل النقاط الطبيعية.

إن للكشف عن الشواذ باستخدام أي من تعريفي الكثافة محاسن ومساوئ مشابهة لتلك الخاصة بمخططات الكشف عن الشواذ استناداً إلى القرابة التي تحدثنا عنها في المقطع 3.10. وبشكل خاص فإنه لا يمكنها تحديد الكائنات الشاذة بشكل صحيح عندما تحوي البيانات مناطق ذات كثافات مختلفة. (انظر الشكل 7.10). ولكي يتم تحديد الكائنات الشاذة في مجموعات بيانات كهذه بشكل صحيح سنحتاج إلى وضع فكرة عامة عن الكثافة نسبة إلى جوار الكائن. النقطة D في الشكل 7.10 على سبيل المثال لها كثافة مطلقة (وفقاً للتعريفين 6.10 و 7.10) أعلى من النقطة A ، ولكن كثافتها أقل نسبة إلى الجيران الأقرب.

توجد طرق كثيرة لتحديد الكثافة النسبية لكائن. إحدى هذه الطرق هي التي تستخدمها خوارزمية العنقدة استناداً إلى الكثافة SNN والتي شرحناها في المقطع 8.4.9. هناك طريقة أخرى وهي حساب الكثافة النسبية كنسبة (ratio) من كثافة نقطة $\mathbf{x}$ والمتوسط الحسابي لكثافة جيرانها الأقرب $\mathbf{y}$ كما يلي:

$$\text{average relative density}(\mathbf{x},k)=\frac{density(\mathbf{x},k)}{\sum_{\mathbf{y}\in N(\mathbf{x},k)} density(y,k)/|N(\mathbf{x},k)|} \quad (7.10)$$

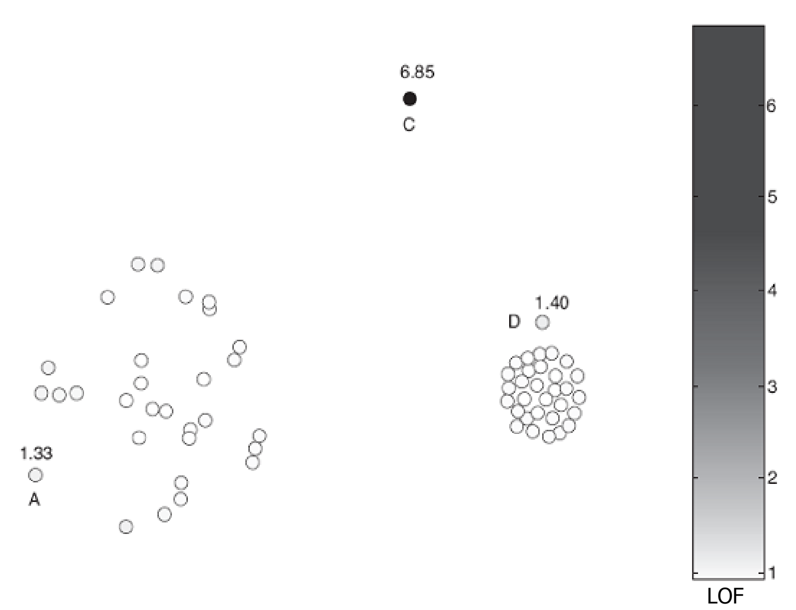
1.4.10 الكشف عن الشواذ باستخدام الكثافة النسبية

سنشرح في هذا المقطع تقنية تستند إلى فكرة الكثافة النسبية. هذه التقنية (وهي نسخة مبسّطة عن تقنية معامل الشذوذ المحلي LOF (Local Outlier Factor)، المشروحة في الخوارزمية 2.10. سنتحدث بالتفصيل عن هذه الخوارزمية فيما يلي، ولكنها باختصار تعمل كما يلي. نقوم بحساب درجة شذوذ كل كائن من أجل عدد محدد من الجيران (k) بأن نحسب أولاً كثافة كائن *density*(**x**, k) استناداً إلى جيرانه الأقرب. يتم بعدها حساب المتوسط الحسابي لكثافة جيران نقطة واستخدامه لحساب المتوسط الحسابي للكثافة النسبية للنقطة كما أشرنا في المعادلة 7.10. يقدم هذا المقدار الكمي دلالة على ما إذا كان **x** موجوداً في منطقة أكثر كثافة أو أكثر تبعثراً للجوار مقارنة بجيرانه ويتم اعتباره درجة شذوذ **x**.

الخوارزمية 2.10 خوارزمية حساب درجة شذوذ استناداً إلى الكثافة النسبية

```
{k is the number of nearest neighbors}
for all objects x do
    Determine N(x, k), the k- nearest neighbors of x.
    Determine density(x, k), the density of x using its nearest neighbors, i.e., the
    objects in N(x, k).
end for
for all objects x do
    Set the outlier score(x, k) = average relative density(x, k) from Equation 10.7.
end for
```

المثال 2.10 (الكشف عن الكائنات الشاذة استناداً إلى الكثافة النسبية). قمنا سابقاً بتوضيح أداء طريقة الكشف عن الكائنات الشاذة استناداً إلى الكثافة النسبية باستخدام مجموعة بيانات المثال الواردة في الشكل 7.10. وبالتالي فإن $k = 10$. يعرض الشكل 8.10 درجات الشذوذ من أجل هذه النقاط. تم تحديد تظليل كل نقطة وفقاً لدرجة شذوذها، بمعنى أن النقاط التي لها درجة أعلى تكون أكثر دُكنة. قمنا بتسمية النقاط A و B و C التي لها أعلى درجات شذوذ باستخدام هذه القيم. إن هذه النقاط بالترتيب هي النقطة الشاذة الأكثر تطرفاً، والنقطة الأكثر تطرفاً بالنسبة لمجموعة النقاط المكتنزة (compact)، والنقطة الأكثر تطرفاً في مجموعة النقاط المفككة. ■



الشكل 8.10 درجات الشذوذ استناداً إلى الكثافة النسبية (LOF) من أجل النقاط ثنائية الأبعاد الواردة في الشكل 7.10

2.4.10 أوجه القوة والضعف

يُعطي الكشف عن الشواذ استناداً إلى الكثافة النسبية مقياساً كميّاً لدرجة كون كائن شاذاً ويمكن أن تعمل بشكل جيد حتى لو كانت البيانات تتضمن مناطق مختلفة الكثافة. وكما هو حال الطرق التي تستند إلى المسافة فإن لهذه الطرق تعقيداً زمنياً قدره $O(m^2)$ (حيث m هو عدد الكائنات)، على الرغم من أنه يمكن تخفيض هذا التعقيد إلى $O(m \log m)$ من أجل بيانات منخفضة الأبعاد باستخدام بُنى بيانات خاصة. من الممكن أيضاً أن يكون اختيار الوسيط صعباً على الرغم من أن خوارزمية LOF المعيارية تعالج هذه المشكلة بالبحث في تشكيلة من قيم k ثم أخذ درجات الشذوذ القصوى. ولكن على أية حال ما زال من الضروري اختيار الحدين الأعلى والأسفل لهذه القيم.

5.10 تقنيات تستند إلى العنقدة

يتم من خلال تحليل العنقدة إيجاد كائنات مرتبطة ببعضها بقوة، في حين أن يتم في الكشف عن التشوهات إيجاد الكائنات التي لا ترتبط بقوة بالكائنات الأخرى. لن يدهشك على هذا

الأساس أن تعلم أنه يمكن استخدام العنقدة للكشف عن الكائنات الشاذة. سنتحدث في هذا المقطع عن عدة تقنيات كهذه.

يتم في إحدى طرق الكشف عن الشواذ باستخدام العنقدة استبعاد العناقيد الصغيرة التي تكون بعيدة جداً عن العناقيد الأخرى. يمكن استخدام هذه الطريقة مع أية تقنية عنقدة، ولكنها تتطلب حدود عتبة من أجل الحجم الأصغري للعنقود والمسافة بين عنقود صغير وبين العناقيد الأخرى. يتم غالباً تبسيط العملية باستبعاد كافة العناقيد التي تكون أصغر من حجم أصغري. إن هذا المخطط حساس جداً لعدد العناقيد المختار. كما أن من الصعب ربط درجة شذوذ بالكائنات باستخدام هذا المخطط. لاحظ أن اعتبار مجموعات من الكائنات شاذة يوسّع (extend) فكرة الشواذ المأخوذة من كائنات منفصلة إلى مجموعات من الكائنات، ولكنه لا يغير أياً من الأساسيات.

هناك طريقة أكثر تنظيماً تتم وفقها أولاً عنقدة كافة الكائنات ثم تحديد درجة انتماء كائن إلى أي عنقود. فمن أجل عنقدة تستند إلى نموذج الأصل (prototype-based)، يمكننا استخدام بُعد كائن عن مركز عنقوده لقياس درجة انتماء كائن إلى عنقود. وبشكل أكثر عمومية، ومن أجل تقنيات عنقدة تستند إلى تابع هدف (موضوعي) (objective)، يمكننا استخدام التابع الهدف لتحديد مدى جودة انتماء كائن إلى أي عنقود. وبشكل خاص، إذا كان ينتج عن حذف كائن تحسّن كبير في قيمة التابع الهدف، فإننا سنقوم بتصنيف الكائن على أنه كائن شاذ. للتوضيح نقول بأنه من أجل K-means يؤدي التخلص من كائن بعيد عن مركز عنقوده إلى تحسن كبير في مجموع مربعات الأخطاء (SSE) للعنقود. وبشكل مختصر نقول أن العنقدة تنشئ نموذجاً من البيانات والتشوهات التي تُحرّف ذلك النموذج. يعبر التعريف 8.10 عن هذه الفكرة.

التعريف 8.10 (الكائن الشاذ استناداً إلى العنقدة). يكون كائن شاذاً استناداً إلى العنقدة إذا كان الكائن لا ينتمي بقوة إلى أي عنقود.

يعتبر هذا التعريف عند استخدامه بواسطة مخططات العنقدة التي لها تابع هدف (موضوعي) حالة خاصة من الكشف عن التشوهات استناداً إلى النموذج. وعلى الرغم من أن التعريف 8.10 أكثر ملاءمة من أجل مخططات تستند إلى نموذج الأصل أو المخططات التي لها تابع هدف، فإنه يشمل أيضاً طرق العنقدة استناداً إلى الكثافة والصلة البينية (connectivity) للكشف عن الشواذ. وبشكل خاص فإنه من أجل العنقدة استناداً إلى الكثافة يكون كائن ما لا

ينتمي بقوة إلى أي عنقود إذا كانت كثافته منخفضة جداً، في حين أنه في العنقدة استناداً إلى الصلة البينية يكون كائن ما لا ينتمي بقوة إلى أي عنقود إذا لم يكن متصلاً بقوة.

سنتحدث فيما يلي عن مسائل تجب معالجتها من قِبَل أي تقنية كشف عن الشواذ استناداً إلى العنقدة. سنركز في حديثنا على تقنيات العنقدة استناداً إلى نموذج الأصل، مثل K-means.

1.5.10 تقييم مدى انتماء كائن إلى عنقود

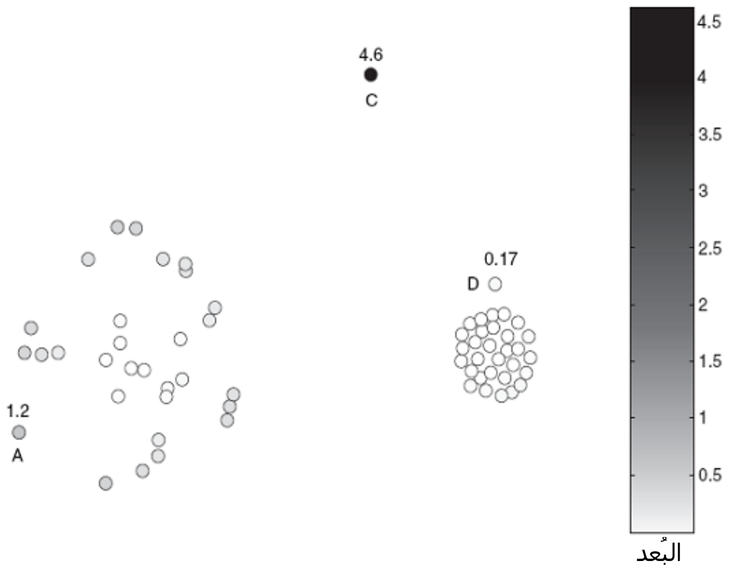
هناك في العناقيد التي تستند إلى نموذج الأصل عدة طرق لتقييم مدى انتماء كائن إلى عنقود. إحدى هذه الطرق هي قياس البُعد بين الكائن وبين نموذج الأصل للعنقود وأخذ هذه المسافة كدرجة شذوذ (outlier score) الكائن. فإذا كان للعناقيد كثافات مختلفة فيمكننا وضع درجة شذوذ تقيس البُعد النسبي لكائن عن نموذج الأصل للعنقود مقارنة بأبعاد الكائنات الأخرى في العنقود. هناك احتمال آخر (إذا كان يمكن نمذجة العناقيد من خلال توزيعات غوصية) هو استخدام مسافة Mahalanobis.

أما بالنسبة لتقنيات العنقدة التي لها تابع هدف فيمكننا إسناد درجة شذوذ إلى كائن تعكس التحسن في التابع الهدف عند حذف ذلك الكائن. يمكن على أية حال أن يكون تحديد درجة كون نقطة ما نقطة شاذة استناداً إلى التابع الهدف أمراً مكلفاً من الناحية الحسابية. ولهذا السبب فإننا نفضّل غالباً الطرق التي تستند إلى المسافة الواردة في الفقرة السابقة.

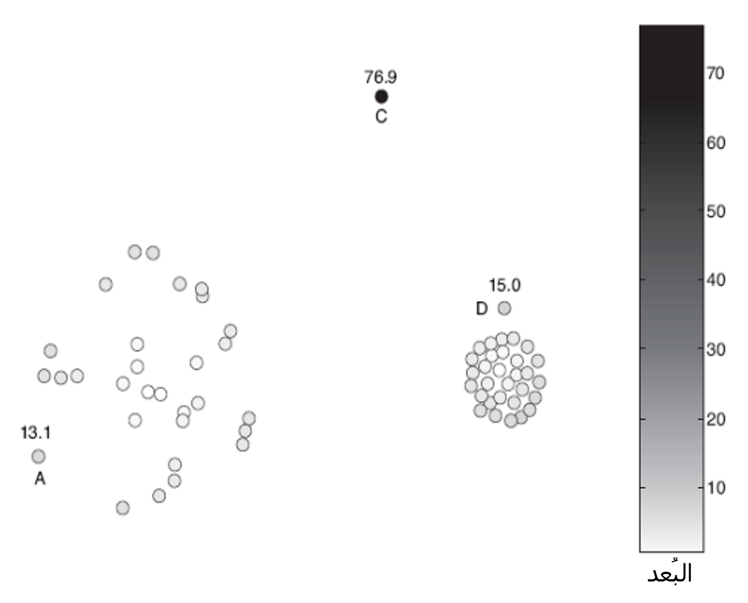
المثال 3.10 (مثال يستند إلى العنقدة). يستند هذا المثال إلى مجموعة النقاط الواردة في الشكل 7.10. تستخدم العنقدة استناداً إلى نموذج الأصل الخوارزمية K-means، ويتم حساب درجة الشذوذ لنقطة بطريقتين: (1) من خلال بُعد النقطة عن أقرب مركز ثقل (centroid) إليها، و (2) من خلال البُعد النسبي لنقطة عن أقرب مركز ثقل إليها، حيث أن البُعد (المسافة) النسبي هو نسبة بُعد النقطة عن مركز الثقل إلى قيمة المسافة الوسط (median distance) لكافة النقاط في العنقود عن مركز الثقل. تستخدم الطريقة الأخيرة للتكيف بحسب الفروقات الكبيرة في الكثافة بين العناقيد الكثيفة والمفككة.

تظهر درجات الشذوذ الناتجة في الشكلين 9.10 و 10.10. وكما في السابق فإننا نشير إلى درجة الشذوذ (تم قياسها في هذه الحالة باستخدام البُعد أو البُعد النسبي) من خلال التظليل. استخدمنا عنقودين في كل حالة. تعاني الطريقة التي تستند إلى المسافة فقط من مشكلات في حالة كون العناقيد مختلفة

الكثافة ، فمثلاً D لا تعتبر شذوذاً. أما بالنسبة للطريقة التي تستند إلى الأبعاد النسبية ، فإن النقاط التي تم اعتبارها شاذة باستخدام LOF (أي A و C و D) قد تبين أنها شاذة هنا أيضاً. ■



الشكل 9.10 بُعد النقاط عن أقرب مركز ثقل



الشكل 10.10 البُعد النسبي للنقاط عن أقرب مركز ثقل

2.5.10 تأثير الشواذ على العنقدة الأولية

إذا تم الكشف عن الشواذ من خلال العنقدة فإن هناك سؤالاً حول ما إذا كانت النتائج مقبولة باعتبار أن الشواذ تؤثر على العنقدة. يمكن استخدام الطريقة التالية للتغلب على هذه المشكلة: تتم عنقدة الكائنات، ثم إزالة الكائنات الشاذة، وبعدها تتم عنقدة الكائنات مرة أخرى. على الرغم من أنه لا توجد ضمانات بأن هذه الطريقة ستعطي نتائج أمثلية، فإن من السهل استخدامها. هناك طريقة أكثر تعقيداً وهي أن تكون هناك مجموعة خاصة من أجل الكائنات التي لا تتلاءم حالياً بشكل جيد في أي عنقود. تمثل هذه المجموعة الكائنات الشاذة المحتملة. ومع تقدم عملية العنقدة فإن العناقيد تتغير. تتم إضافة الكائنات التي لم تعد تنتمي بقوة إلى أي عنقود إلى مجموعة الكائنات الشاذة المحتملة، في حين يتم فحص الكائنات الموجودة حالياً في المجموعة لنرى ما إذا كانت تنتمي الآن بقوة إلى عنقود ويمكن إزالتها من مجموعة الكائنات الشاذة المحتملة. يتم تصنيف الكائنات التي تبقى في المجموعة في نهاية العنقدة على أنها شاذة. وهنا أيضاً لا توجد ضمانات بأن نحصل على حل أمثلي أو حتى ضمانات بأن هذه الطريقة ستعمل بشكل أفضل من الطريقة الأبسط المشروحة سابقاً. فمثلاً، قد يبدو عنقود يحوي نقاط تشويش (noise) مشابهاً لعنقود فعلي بدون شواذ. يمكن أن تكون هذه المشكلة جدية خاصة إذا كان يتم حساب درجة الشذوذ باستخدام المسافة (البُعد) النسبية.

3.5.10 عدد العناقيد المستخدمة

لا تحدد تقنيات عنقدة مثل K-means بشكل أوتوماتيكي عدد العناقيد. وهذه تعتبر مشكلة عند استخدام العنقدة في الكشف عن الشواذ، باعتبار أن تحديد كون كائن ما شاذاً أم لا يعتمد على عدد العناقيد. فمثلاً، قد تكون مجموعة من 10 كائنات قريبة نسبياً من بعضها، ولكنها قد تكون محتواة كجزء من عنقود أكبر إذا وجدنا بضعة عناقيد كبيرة. وفي هذه الحالة فإن من الممكن اعتبار النقاط العشر شاذة، حتى بالرغم من أنها كانت ستشكل عنقوداً فيما لو تم تحديد عدد أكبر للعناقيد.

لا يوجد حل بسيط لهذه المشكلة، كما هو حال بعض القضايا الأخرى. يمكن مثلاً تكرار التحليل من أجل عدد مختلف من العناقيد. كما يمكن إيجاد عدد كبير من العناقيد الصغيرة. الفكرة هنا هي أن (1) العناقيد الصغيرة تميل إلى أن تكون أكثر تماسكاً، و (2) إذا كان كائن شاذاً حتى عندما يكون هناك عدد كبير من العناقيد الصغيرة، فإنه على الأغلب سيكون شاذاً فعلاً. إلا أن المشكلة هي أن مجموعات الشواذ تلك قد تشكّل عناقيد صغيرة وبالتالي لن يتم كشفها.

4.5.10 أوجه القوة والضعف

تمتلك بعض تقنيات العنقدة (مثل K-means) تعقيداً خطياً (linear) أو غير خطي من حيث الزمن أو المساحة، وبالتالي من الممكن أن تكون تقنيات الكشف عن الشواذ استناداً إلى خوارزميات كهذه عالية الفعالية. كما أن تحديد عنقود هو أمر متمم لتحديد الشواذ، وبذلك فإن من الممكن عادة إيجاد العناقيد والشواذ معاً في نفس الوقت. أما من الناحية السلبية فإن مجموعة الشواذ الناتجة ودرجات شذوذها يمكن أن تعتمد بشكل كبير على عدد العناقيد المستخدم بالإضافة إلى وجود شواذ في البيانات. فمثلاً، يمكن أن يحرف وجود الشواذ العناقيد الناتجة عن خوارزميات تستند إلى نموذج الأصل. تتأثر جودة الكائنات الشاذة الناتجة عن العنقدة بجودة العناقيد التي تعطيها الخوارزمية. وكما قلنا في الفصلين 8 و 9 فإن كل خوارزمية عنقدة تكون مناسبة فقط من أجل نوع معين من البيانات، وبالتالي فإن من الضروري اختيار خوارزمية العنقدة بعناية.

6.10 التمارين

1. قارن بين التقنيات المختلفة للكشف عن التشوهات الواردة في المقطع 2.1.10. وبشكل خاص، حاول تحديد الظروف التي تكون فيها تعريفات الكائنات المشوّهة المستخدمة في التقنيات المختلفة غير متكافئة أو الحالات التي يكون فيها تعريف ذا معنى فيما يكون تعريف آخر لا معنى له. لا تنسَ أن تأخذ بعين الاعتبار الأنواع المختلفة للبيانات.

2. يمكن استخدام تحليل الاقتران لإيجاد التشوهات كما يلي. يتم إيجاد أنماط اقتران قوية تشتمل على عدد أصغري ما من الكائنات. الكائنات المشوهة هي تلك الكائنات التي لا تنتمي إلى أي نمط كهذا. لفهم ذلك بشكل أفضل، لاحظ أن نمط اقتران العصبة المتشعبة (hyperclique) الذي تحدثنا عنه في المقطع 8.6 مناسب بشكل خاص لهذه الطريقة. وبشكل خاص، إذا كان لدينا مستوى h-confidence يحدده المستخدم، فسيتم إيجاد أنماط عصبة متشعبة أعظمية (maximal hyperclique pattern). يتم تصنيف كافة الكائنات التي لا تظهر في أنماط العصبة المتشعبة الأعظمية على أنها شواذ.

3. اشرح التعقيد الزمني المحتمل لطرق الكشف عن التشوهات التي تستند إلى الطرق التالية: استناداً إلى النموذج (model-based) باستخدام العنقدة، واستناداً إلى القرابة، والكثافة. لا

نحتاج إلى معرفة بتقنيات محددة. ركز نوعاً ما على المتطلبات الحسابية الأساسية لكل طريقة، مثل الزمن المطلوب لحساب كثافة كل كائن.

4. يعتبر اختبار Grubb (المشروح من خلال الخوارزمية 3.10) إجرائية أكثر تعقيداً من الناحية الإحصائية للكشف عن الشواذ مقارنة بالتعريف 3.10. هذه الإجرائية تكرارية وتأخذ أيضاً بعين الاعتبار حقيقة أن قيمة z ‏(z-score) لكل قيمة تستند إلى متوسط والانحراف المعياري للعينة من أجل مجموعة القيم الحالية. يتم استبعاد القيمة التي لها أعلى قيمة z إذا كانت قيمة z لها أكبر من g_c، وهي القيمة الحدية (critical) للاختبار من أجل تحديد كائن شاذ عند مستوى أهمية (significance level) هو α. يتم تكرار هذه العملية إلى أن لا تكون هناك كائنات يتم استبعادها. لاحظ أنه يتم تحديث قيم متوسط والانحراف المعياري للعينة و g_c عند كل تكرار.

الخوارزمية 3.10 طريقة Grubb لاستبعاد الكائنات الشاذة

Input the values and α

{m is number of values, α is a parameter, and t_c is a value chosen so that $\alpha = prob(x \geq t_c$ for a t distribution with m-2 degrees of freedom.}

repeat

Compute the sample mean ($\bar{x}$) and standard deviation (s_x).

Compute a value g_c so that $prob(|z| \geq g_c) = \alpha$

(In terms of t_c and m, $g_c = \frac{m-1}{\sqrt{m}}\sqrt{\frac{t_c^2}{m-2+t_c^2}}$.)

Compute the z-score of each value, i.e., $z = (x-\bar{x})/s_x$

Let $g = \max|z|$, i.e., find the z-score of largest magnitude and call it g.

if $g > g_c$ **then**

Eliminate the value corresponding to g.

$m \leftarrow m - 1$

end if

until No objects eliminated.

(a) ما هو الحد الأقصى للقيمة $\frac{m-1}{\sqrt{m}}\sqrt{\frac{t_c^2}{m-2+t_c^2}}$ المستخدمة في اختبار Grubb عندما تقترب m من اللانهاية؟ استخدم مستوى أهمية 0.05.

(b) اشرح (بالكلمات) معنى النتيجة السابقة.

5. تُعطى الكثافة الاحتمالية لنقطة **x** وفقاً لتوزيع طبيعي متعدد المتحولات (المتغيرات) له متوسط μ ومصفوفة تباين مشترك Σ من خلال المعادلة التالية:

$$prob(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^m |\sum|^{1/2}} e^{-\frac{(\mathbf{x}-\mu)\sum^{-1}(\mathbf{x}-\mu)}{2}} \qquad (8.10)$$

فإذا استخدمنا متوسط العينة $\bar{\mathbf{x}}$ ومصفوفة التباين المشترك **S** كتقديرين للمتوسط μ ومصفوفة التباين المشترك Σ (على الترتيب)، فبيّن أن log *prob*(**x**) يساوي مسافة Mahalanobis بين نقطة بيانات **x** ومتوسط العينة $\bar{\mathbf{x}}$ مضافاً إليه ثابت لا يعتمد على **x**.

6. إذا كان لدينا مخطط K-means (المسافة النسبية) للكشف عن الكائنات الشاذة المشروح في المقطع 5.10 والشكل المرافق (الشكل 10.10).

(a) إن للنقاط الموجودة في أسفل العنقود الكثيف الذي يظهر في الشكل 10.10 درجات شذوذ عالية نوعاً ما مقارنة بتلك النقاط الموجودة في أعلى العنقود الكثيف. لماذا؟

(b) لنفترض أننا اخترنا عدد العناقيد بحيث يكون أكبر، كأن يكون 10 مثلاً. هل ستبقى التقنية المقترحة فعالة في إيجاد الكائنات الشاذة الأكثر تطرفاً الموجودة في أعلى الشكل؟ لمَ، أو لمَ لا؟

(c) يؤدي استخدام المسافة (البُعد) النسبية إلى التكيّف بحسب فروقات الكثافة. أعط مثالاً عن حالة تقودنا فيها هذه الطريقة إلى النتيجة الخطأ.

7. إذا كان احتمال أن يتم تصنيف كائن طبيعي على أنه تشوه هو 0.01 واحتمال أن يتم تصنيف كائن مشوّه على أنه تشوّه هو 0.99، فما هي نسبة الإنذار الخاطئ (false alarm rate) ونسبة الكشف (detection rate) إذا كان 99% من الكائنات طبيعية؟ (استخدم التعريف الوارد أدناه).

$$\text{detection rate} = \frac{\text{number of anomalies detected}}{\text{total number of anomalies}} \quad (9.10)$$

$$\text{false alarm rate} = \frac{\text{number of false anomalies}}{\text{number of objects classified as anomalies}} \quad (10.10)$$

8. بفرض أن لدينا مجموعة من النقاط، حيث أن معظم هذه النقاط موجود في مناطق منخفضة الكثافة، ولكن هناك بضعة نقاط موجودة في مناطق ذات كثافة عالية. فإذا قمنا بتعريف الكائن المشوه على أنه نقطة في منطقة منخفضة الكثافة، فسيتم تصنيف معظم النقاط على أنها تشوهات. هل تعتبر هذا استخداماً مناسباً لتعريف التشوه الذي يستند إلى الكثافة أم يجب تعديل التعريف بطريقة ما؟

9. لتكن لدينا مجموعة نقاط تخضع للتوزيع المنتظم ضمن المجال [0, 1]. هل تعتبر أن الفكرة العامة القائلة بأن الكائن الشاذ هو قيمة لا تتم مشاهدتها بشكل متكرر معبرة من أجل هذه البيانات؟

10. قام محلِّل بتطبيق خوارزمية كشف عن التشوهات على مجموعة بيانات ووجد مجموعة من التشوهات. إلا أن فضوله قد دفعه لتطبيق خوارزمية الكشف عن التشوهات على مجموعة التشوهات.

(a) ناقش سلوك كل واحدة من تقنيات الكشف عن التشوهات المشروحة في هذا الفصل. (وإذا كان ذلك ممكناً، قم بتجربة هذا الأمر على مجموعات بيانات وخوارزميات فعلية).

(b) ما السلوك الذي تعتقد أن خوارزمية الكشف عن التشوهات ستسلكه عند تطبيقها على مجموعة من الكائنات المشوّهة؟